Review Team #2

A Statistical Review of 'Developing Methods to Reduce Bird Mortality in the Altamont Pass Wind Resource Area' by Smallwood and Thelander (August 2005, Publication #500-04-052)

Original draft: **June 26, 2006**
Revised draft: September 11, 2006

Commissioned by:

The California Institute for Energy and Environment on behalf of the California Energy Commission

Accessing the report:

The Smallwood/Thelander report can be downloaded from the following web site: www.energy.ca.gov/pier/final_project_reports/500-04-052.html

Comments regarding revision of the original June 26, 2006 peer review:

The original version of this review was submitted on June 26, 2006. On August 29, 2006, the three groups of reviewers received the Smallwood and Thelander responses to their original review. Responses to the Smallwood and Thelander comments to this review are addressed in two ways. Additions, such as this, are italicized, and deletions will be indicated in footnotes. Be aware, however, that the original review did contain a few italicized phrases. In the footnote, the deleted parts will have ~~strikethrough~~ lines through them.

Purpose of the review:

The purpose of this anonymous review is to objectively evaluate the statistical analysis used in the report 'Developing Methods to Reduce Bird Mortality in the Altamont Pass Wind Resource Area' by Smallwood and Thelander (August 2005, Publication #500-04-052) created for the California Energy Commission (CEC). More specifically, the reviewers will evaluate whether the data collection and statistical analysis methods are scientifically sound and appropriate for achieving the report goals set forth by the CEC. Policy recommendations are not to be reviewed.

This review was created by "Review Team 2". Review Team 2 was comprised of three individuals working together. One member's professional training is as a biostatistician, another as a wildlife ecologist, and the third as an environmental engineer.

Abbreviations and notation:

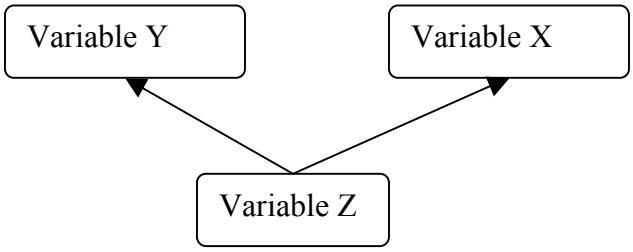
Certain abbreviations and notation will be used in this review:

- (p.73, par.2) = Page 73, paragraph 2. The first paragraph is considered the text at the top of the page, regardless of whether the text's paragraph began on that page or the previous page.
- Altamont Pass Wind Resource Area = APWRA

Terminology commonly used:

When evaluating a report based upon data analysis, understanding certain statistical terminology and ideas are necessary. We hope the following explanations will help the readers of this review better understand some basic statistical concepts.

- **Confounding:** Confounding is statistical terminology for when the researcher cannot tell if variable Y is caused by variable X or variable Z. This confusion arises because variables X and Z are related. Sometimes a dataset only has the X and Y variables and the analysis statistically indicates that Y and X are associated. It may be, however, that Y is caused by the unmeasured Z variable and that Z also causes X; X does not cause Y. Consequently the word "associated" is often used in statistics because variable X may not cause Y but it is associated with Y. (See the following figure.)



- **Null and Alternative Hypotheses:** The null hypothesis is considered to be the status quo and is to be retained by the researchers unless the data suggest strongly otherwise. (Much like the saying, "Innocent until proven guilty beyond a reasonable doubt.") The alternative hypothesis is often considered the "research hypothesis" and is in contrast to the null hypothesis. For example, a null hypothesis would be that wind turbine model A has the same bird mortality rate as model B. The alternative hypothesis, which must be stated before observing the data, could be that turbine model A has a lower bird mortality rate compared to model B. Or the alternative hypothesis may be that the mortality rates for the two models are different. If the probability of getting the collected data is considered to be suspiciously improbable assuming the null hypothesis is true and the collected data would have been more probable if the alternative hypothesis is true, the null hypothesis will be rejected in favor of the alternative hypothesis. This finding is considered "statistically significant".
- ***P*-value:** This is a probability that states the likelihood of getting the sampled data, or data even more unlikely, if the null hypothesis were true. Statistical calculations are performed assuming the null hypothesis is true. Typically, if the *P*-value is equal to or less than 0.05, there is considered to be enough doubt to believe that the data were not generated with the null hypothesis being true, but instead the data were generated with the alternative hypothesis being true. It is conventional to consider the findings to be "statistically significant" if the *P*-value is less than or equal to 0.05. Statistical significance is a measure of probability concerning the null hypothesis, not of the magnitude of the effect being investigated.
- **Type I error:** This error is often called a "false negative" or a "rejection error". This occurs when the null hypothesis is rejected in favor of the alternative hypothesis even though (unknowingly) the null hypothesis was true. If using a cut-off of 0.05 for the *P*-value, for every 20 statistical tests performed when the null hypothesis is actually true we would expect to commit one Type I error.
- **Type II error:** This error is often called a "false positive" or an "acceptance error". This occurs when the null hypothesis is retained although the alternative hypothesis is true. The probability of this event occurring depends upon many factors.
- **Power:** The probability of rejecting the null hypothesis when the alternative hypothesis is true is called the power of the test. This is the flip-side of a Type II error and can be thought of as the sensitivity of a statistical test – the likelihood of

giving a true positive. It depends upon many factors, but primarily the magnitude of the effect (if it exists) being tested for, the amount of natural and measurement variation in the variable being measured, and the sample size.

Report Overview:

The authors were challenged with a broad question, a large geographic region, limited access, and a very large number of wind turbines of various models. Their report is a quantitative exploration into the variables associated with bird mortality. Almost certainly, at least some of the many variables measured are truly linked to bird mortality – birds are certainly being killed in the APWRA. The reviewers have little confidence, however, that this report has scientifically been able to determine which of those variables are important.

Much effort went into collecting massive amounts of data; however, the authors should have focused more effort on study design and collected their data more wisely. Likewise, the data analyses could have been more thoughtful and sophisticated. The statistical analyses are applied in an automated manner that fails to fully utilize the data at hand and ignores potential confounding of variables. It seems like many of the statistics were calculated just for the purpose of producing statistical tables to the point of data dredging. Furthermore, the mathematical assumptions behind statistical tests like one-way ANOVA are ignored and thus the reported P-values should be treated as approximations. The large number of statistical tests likely resulted in many Type I errors; therefore, statistically significant findings should be treated more as an indicator of what should be explored in future studies.

Furthermore, the reviewers concur with the authors that their study does not give accurate estimates of actual bird mortality.

Broader comments addressing specific questions:

- *Was the statistical methodology used on the analysis consistent with accepted methods used in other biostatistical analyses?*

No. A very large number (>1000) of univariate chi-square tests is not common in biostatistical analyses. Interpretations of the univariate tests are clouded somewhat by shared variation among the explanatory variables (turbine attributes).

Chi-square analysis assumes that the counts are exact and not estimated counts[1]. *Adjusted counts (adjusted for scavenging and detection rates) are frequently used throughout the report and it is not always obvious whether adjusted or raw counts are being used to calculate the statistics. If adjusted counts, rather than raw counts, are used in a chi-square test, it* is not clear how *the uncertainty in adjusted counts* would influence the conclusions reached on the numerous chi-square hypothesis tests.

[1] Original review text before considering the Smallwood and Thelander response: "Chi-square analysis assumes that the counts are exact and not estimated counts ~~*as* they are in this study~~. It is not clear how ~~this~~ would influence the conclusions reached on the numerous chi-square hypothesis tests."

In estimating mortality rates for specific species due to wind turbine collisions, almost half (28) of the 60 species or groups have fewer than 5 fatalities reported in the entire project. And yet, mortality rates are still estimated and reported.

Although the study design is observational, the authors quickly jump to *confirmatory analysis methods such as* hypothesis testing and parametric analysis without exploring their datasets thoroughly. *The application of the exploratory data analysis methods such as those developed and popularized by John W. Tukey, Frederick Mosteller, and others would have been more appropriate.* What distinguishes the 20% of the turbines where fatalities were discovered from the 80% without fatalities?

Most biostatistical analyses based on analogous (but often smaller) datasets rely on multivariate models such as a general linear models, logistic or Poisson regressions, or discriminant function analyses. The authors explain that limitations in their sampling precluded these more sophisticated multivariate analyses, but this may not be true if the authors (a) carefully screen their variables to reduce the number of parameters in their models, and/or (b) clearly restrict their inferences to the turbines actually sampled.

- *Were the technical approaches used in the research appropriate for achieving stated goals?*

The stated goals for this study were to 1) quantify bird use, 2) evaluate the flying behaviors and conditions associated with flight behaviors, 3) identify the relationships between bird mortality and various explanatory variables, and 4) develop predictive, empirical models that identify areas or conditions associated with high vulnerability.

The sampling programs designed to address the first goal are not best for goals 3 and 4. (The authors used a separate approach for goal 2, which unfortunately omitted the summer.) Consequently the study design is not well crafted for achieving objectives 3 and 4.

- *Were the data collection and analysis methods and assumptions clearly stated, valid, and reliable? Were there any errors or, flaws? Were any relevant factors missing?*

The authors used *recently developed* protocols for carcass searches, *and standard* bird observations, rodent surveys, etc. to obtain the ecological data, and the technical approaches were appropriate.[2] However, the methods used to estimate bird mortality rate are suspect because (a) neither scavenging rate nor observer detection probabilities were measured empirically, values were pulled from the literature – in

[2] Original review text before considering the Smallwood and Thelander response: "The authors used ~~standard~~ protocols for carcass searches, bird observations, rodent surveys, etc. to obtain the ecological data, and ~~generally~~ the technical approaches were appropriate."

some cases based on studies in different locations; (b) a 50m search radius is insufficient to detect an adequately high percentage of carcasses, especially given the lack of rigorous data on detection rates of carcasses beyond 50m from a tower; (c) the authors adopted adjustments to published scavenging and detection rates based on assumptions that are inadequately supported with observation. For example, the following three assumed adjustments are problematic: (1) "halving" the scavenging rate for raptors, (2) elevating the scavenging rate by 10% for the 2nd set of turbines because they were checked much less frequently than those in the study from which scavenging rates were used, and (3) assuming detection rates were equally high beyond 50m, where the crews did not search rigorously. Most of these inadequacies biased mortality estimates by an unknown amount and direction. For comparative purposes of a single species' mortality rates across turbine and location attributes (Chapter 7), these biases may operate *roughly* similarly across the variables and therefore may not undermine the analysis. For examination of impact (Chapter 4), however, these biases are very problematic indeed.

- *Was the study design scientifically sound? Was there sufficient time to conduct the study (e.g., time for conducting searches, time for assessing seasonal effects)?*

The description of the sampling – how well it yielded a representative sample of all the turbines in APWRA – was inadequate, hindering our ability to rigorously assess the sufficiency of the sampling itself. Access to study the 2nd set of turbines was granted too late and the study's duration (and hence the length of their reexamination) was too short to be of maximum use to the overall project. The bird behavioral sampling did not include most of the summer season.

The sampling design is not clear. What is the sampling element? Is it the turbine or the turbine string or is it the sampling visit? How was the order selected for visiting the strings?

- *Were uncertainties described, either qualitatively or quantitatively?*

In some cases, yes; however, the very large number of univariate test significantly inflates the probability of false positive results across the entire project. The authors made no attempt to adjust, quantify, and describe this issue.

In addition, many estimates of rates were provided with no attempt to describe the associated uncertainties. For example, tables 7-4 through 7-7, 7-9, 7-11, 7-13, and 7-15, all provide estimates of *the percentage*[3] increase in mortality associated with a given variable, but no qualitative or quantitative measures of uncertainty are provided. *A percentage change is a proportion (change in number of fatalities/total fatalities) and confidence intervals for proportions are easily computed for large*

[3] Original review text before considering the Smallwood and Thelander response: "... all provide estimates of ~~rates of~~ increase in mortality..."

sample sizes and also, with somewhat more effort, for small samples. Similarly, the species or group specific mortality rates given in tables 3-11 and 3-12 are presented with no *statistical* measures of uncertainty provided*; i.e., they provide low and high estimates mortality, but do not provide the reader with a statistical measure of the quality of these estimated bounds* .

The authors do not consider any interactions, which further inflates the magnitude of the uncertainties.

- *Were findings statistically significant?*

It is likely that some number of the reported test results were statistically significant. But due to the very large number of univariate tests conducted, there is a high probability that a number of "significant" results were based on pure chance. With an accepted *P*-value of 0.05, then 5 out of every 100 tests will, on average, appear statistically significant by chance when the null hypothesis is true. No effort to account for this was made by the authors.

- *Were the conclusions supported?*

We cannot accept this analysis as one that has rigorously tested hypotheses regarding determinants of bird mortality and that could be reasonably applied in decision making. Instead, it may be more useful to consider this project an exploratory analysis that has identified a number of variables positively associated with increased mortality rates. Therefore, the product of this research is an educated list of working hypotheses. This valuable contribution can be followed by more thorough testing of said hypotheses by rigorous sampling and controlling of confounding variables via sophisticated multivariate analysis of observation data and/or controlled experimentation.

- *Other observations and comments?*

See specific comments, below.

Chapter 1: Understanding the Problem

- p.9, par.4 and 5: The authors imply that a "use vs. availability" approach to quantifying vulnerability can be effectively pursued via chi-squared tests. *A classic*[4] paper describing chi-square (goodness-of-fit) tests to examine use vs. availability of resources in a wildlife context is by Neu et al. (1974). Since that paper was published over 30 years ago, resource selection analyses involving so called use-versus-availability designs have advanced substantively (especially in the last 10 years). Now, few biologists would consider chi-square tests[5] state-of-the-art for use-versus-availability designs (see book on the subject by Manley et al. 2002 and Journal of Wildlife Management volume 2006 issue #2). Instead, most use-versus-availability designs make use some form of logistic regression functions or general linear models. *In fact, Thomas and Taylor (2006, in said volume of J. Wildlife Management), found that 35% of recent use-vs-availability studies use logistic regression; only 8% used chi-square goodness-of-fit tests.*

  Even under the context of using the chi-square test for use-and-availability analysis, the calculations require that the authors have, for each particular bird species, the number killed at turbines in a particular landscape type, the number killed in all landscape types, and the proportion of landscapes that are of that particular landscape type. The chi-square test assumes that the observed counts are accurate and any variation occurs simply from chance and not from observer error. As stated frequently in following chapters, the actual mortality counts are actually *estimated* counts and assumed to be biased low. Even assuming the mortality estimated counts are not biased low or high, this will result in inaccurate levels of statistical significance for the chi-square tests.

  And finally, chi-square tests are typically of two types: test for association/independence and test for goodness-of-fit. These chi-square tests are goodness-of-fit tests where the null hypothesis is that the counts were generated by a uniform distribution. That is, if there were no preference for the various categories of the explanatory variable, a carcass (or whatever response variable is being measured) would be equally likely to end up in any of the categories when adjusted for availability of the categories.

  Manley, B. F. J., L. L. McDonald, D. L. Thomas, T. L. McDonald, and W. P. Erickson. 2002. Resource selection by animals. Second edition. Kluwer Academic Publishers, Dordrecht, Netherlands.
  Neu et al. 1974. A technique for analysis of utilization-availability data. *Journal of Wildlife Management* 38:541-545.

- p.12, par.1: "… we are able to identify which environmental factors might have a causal relationship." After so many years of studying avian mortality associated with wind turbines prior to this work, exploratory observational studies should be

[4]Original review text before considering the Smallwood and Thelander response: "~~The "original"~~ paper ..."
[5]Original review text before considering the Smallwood and Thelander response: "...few biologists would consider chi-square tests ~~effective or~~ as state-of-the-art..."

superseded by designed experimental studies. Observational studies are not able to reveal causality. Experiments, however, can show causality. Yet there is no evidence here that any experimental design took place prior to the observations. The sample locations and times were certainly not random nor were they seemingly selected to provide contrasts in factor levels. This would have allowed them to better compare the variables of interest and help to eliminate confounding variables.

- p.14, Figure 1-1: This is a useful location map; however, a more useful map would have shown the topographic and other specific features of the APWRA. Are there distinct regions of the resource area that might be used to stratify the design?

- p.19, Table 1-1: On p.13, par.3, Table 1-1 is described as "…summarizing the wind turbine attributes of the wind turbines in our sample in the APWA." Much more information is needed here. If this is the sample, how many of each type of turbine is in the sample? How many observations (visits?) occurred at each turbine type in the first set and in the later one? What fraction of the total turbines in the APWRA does each of these types constitute? A description of the sample and the population is called for here. Are these turbines representative of the entire APWRA population?

  Some information is provided in section 7.3.1, but it focuses on sampled turbines attributes and does not provide adequate comparison to the target population (not to mention it is in Chapter 7 on page 189…a long time to wait for readers who will naturally wonder about this issue beginning in Chapter 1). In the end, the study reports data from 4,074 turbines (some with more data than others), and 1,326 turbines remained unmapped and characterized (these numbers were most easily extracted on page 352 in Chapter 9, and in our opinion should be made very prominent here in Chapter 1). But after a complete reading, the reader is still left wondering this most basic of questions – did the sampled turbines adequately represent all the turbines in APWRA? The authors need to provide a table summarizing the distribution of the sampled turbines (both sets) relative to the complete "population" of turbines. We recognize that there may be some variables that the authors cannot ascribe to turbines that were not studied (e.g., grass height surrounding the turbine), but we assume many variables are catalogued by the turbine owners (turbine model, rotor speed, etc.) and/or obtainable from GIS (elevation, slope, aspect, etc.). Figures 1-2 through 1-7 provide visuals of the distribution of sampled turbines, but they offer no information on how these distributions compare to the target population because the unstudied turbines are simply marked "unmapped", *prohibiting a visual comparison of the sampled population versus the target population*. [6]

[6] Original review text before considering the Smallwood and Thelander response: "~~This is a significant shortcoming of the report.~~"

On more minor notes, why are model numbers only given for the Kenetech turbines? The column headed "Size (kW)" should be headed "Rated Power (kW)".

- p.21, Figure 1-2 through Figure 1-7: These are the first of many colorful figures of this type in this report. Each one shows the spatial distribution of some factor. It would be useful to include an additional figure that depicts which turbines were linked to 1 carcass, 2 carcasses, etc.

Chapter 2: Cause of Death and Locations of Bird Carcasses in the APWRA

- It would seem appropriate to present the methods section, given in Chapter 3, prior to reporting the results. It is not possible to make sense out of the various results given in Chapter 2 without knowing the sampling methods used and the underlying sampling program design.

- p.28, par.5: The authors state that one-way analysis of variance (ANOVA) is commonly used and least significant differences (LSD) to compare groups. The authors should give detail as to which LSD method was used as there are several different variations, although it is doubtful this resulted in any significant changes in their calculations.

  A more important defect is the authors' excessive use of one-way ANOVA throughout this chapter and report. Many variables are tested one by one for association with mortality using one-way ANOVA. This approach makes the analyses vulnerable to confounding variables when two are more variables are highly correlated with one another, such as blade height and blade speed. The basic statistical rule that "association is not causation" can get lost in data analysis expeditions. In addition, each time a one-way ANOVA analysis is performed, the data should be graphed so that readers can see if a particular characteristic of the dataset is having heavy influence on the outcome and whether or not more subtle statistical theory violations are occurring. In light of the absence of such graphs, the *P*-values can be considered only approximate at best.

  Given the phenomenal number of univariate hypothesis tests done later in this report, it is surprising that there is no discussing of corrections for multiple comparisons here.

  It would also be helpful if the authors stated which statistical software package was used to do these analyses.

- p.29, par.2: What are the dates for season boundaries? These are not presented until Chapter 7 on page 182. Even there, the description of these dates and why they were chosen is inadequate (see later comments).

  How were days since death estimated? Were these simply guessed via personal experience? How was such experience gained?

- p.32, Table 2-1: Of the 1162 detected birds (and bats) killed by turbine collisions, almost 50% (49.5%) were restricted to 4 of the 60 species/groups reported: Red-tailed Hawk (18.3%), Rock Dove (16.9%), Western Meadowlark (8.3%), and Burrowing Owl (6.0%). Does this high concentration (i.e., 50% of deaths in 7% of the species) reflect the differences in a) abundance among these species, b) the relative risk of wind turbine collisions, or c) the probability of carcass detection?

- p.38, par.1: The methods used to search for carcasses are not described until Chapter 3. This makes the understanding of Chapter 2 material awkward for the readers unless Chapter 3 has already been read.

  The authors openly stated earlier that their search radius was 50 meters (m) and acknowledge that some "unknown proportion" of carcasses outside of the search radius went uncounted (p.28, pars.1 and 2). Yet, an unsupported statement is made here (p.38, par.1) that the "search radius included 84.7% of the carcasses of large-bodied bird species determined to be killed by wind turbines or unknown causes." How was this 84.7% calculated? In light of their search radius, it is not surprising that the majority of the carcasses were found inside the 50m radius of wind turbines. This problem is repeated later (p.42, par.5) when they note that their search radius "included 90.5% of the carcasses of small-bodied bird species." How they determine "90.5%" is left totally unclear to the reader.

  It is unclear both in this section and in Chapter 3 how the carcasses beyond 50m from the turbines were discovered. If the discoveries were accidental and not within the defined sample element, then why were they included in the analysis? If the discoveries beyond 50m were accidental, describe the circumstances of the accidents. Were the observers walking in toward or away from the turbine strings? If they were collected as part of a special study in a systematic search that extended beyond the 50m limit, then describe that study's methods and results.

- p.39, Figures 2-9 and 2-8: These figures confirm that the authors found and counted carcasses found well beyond their 50 meter search radius. That some were found as far as 200 and 220 meters distant make the idea of happenstance discovery of carcasses outside of a systematic search procedure more believable. How were these carcasses found?

If the discoveries shown in these figures beyond 50m were accidental, then, whatever the resultant pattern, it is unreliable since different sampling effort was expended within the 50m limit then beyond it. Consequently, we expect to have more discoveries within 50m then beyond it. It is no surprise that 75% of the large bodied birds were found within 42m of the tower. If we had a uniform density of birds on the ground in a 50m radius of the tower, we would expect to find 74% of the birds within 43m of the tower as shown in this simple ratio circles' areas $\frac{\left(\pi \times 43^2\right)}{\left(\pi \times 50^2\right)} = 0.74$.

Imposing a normal curve on this is unwarranted and somewhat misleading. The only patterns that are worth analyzing are within the 50 m limit. Within that limit, the distributions of discoveries with distance are similar for both large and small bodied birds. As a very minor note from the reviewers, applying the normal distribution curve to these bar graphs is not sensible considering the truncation at 0 meters and that the first bar represents only a 5 meter range while the other bars cover 10 meters. This is likely an artifact of the statistical software, but can be specified by the users. Later, the authors also put the normal curve into bar plots for non-random variables which are determined by the authors such as number of searches (Figure 3-1, p.49).

- p.40, Figure 2-10: A polar or wind rose plot would be clearer. How can the 0 and 360 degrees cells not have identical counts since they are the same direction? What is the predominant wind direction? And what about the direction the wind turbine is facing?

- p.41, Figure 2-11 and referring text p.38, par.2 and p.42, par.6: The authors use simple linear regression to show that mortality counts increase linearly with turbine tower height. The mathematical assumptions behind linear regression are not valid with this particular dataset (likely nonlinearity, non-normal distribution of errors, unequal variances) thus inadequately demonstrating statistically conclusive evidence that mortality counts are greater for taller turbines. The fact that the one-way ANOVA for wind turbine model and carcass distance was statistically insignificant (p.42, par.7) suggests the height-distance conclusion is questionable. In a confused sequence of logic, the authors state (p.42, par.6), "[the regression] predicted that for every meter increase in tower height, average distance of the carcass from the tower increased by half a meter." This clearly ignores that different wind turbine models have different tower heights, thus it may not be the height, but rather the model, that results in the carcass distance. Height and wind turbine model are confounding variables.

  The authors stated, "Distance from tower [to the carcasses] increased with tower height, according to regression analysis, although the precision was poor." The overwhelming majority of the towers were 18.5m and 24-25m tall, making this

primarily a study of these towers with a few others added in. Consequently, the observations at the lowest and highest towers had the greatest influence on the regression. [7] Even with the data for the 43m towers, the regressions only explain a trivial 1% of the variance in the distances that the carcasses were found from towers. The phrase "poor precision" is an understatement. This is the difference between "statistically significant" and biologically important.

- pp.42 and 44: A description of the tower population would be useful here. For the sampled towers and the population as a whole, how many towers of each type, what elevation distribution, what string lengths (1 to n), what spacing between towers in string, etc?

  The authors survey how carcass distance relates to multiple independent variables including tower height (continuous); blade speed (continuous); upwind vs. downwind (binomial); end, gap, or interior of string (categorical); season (categorical); whether turbine was in a canyon (categorical), slope grade (categorical); or elevation (continuous). They investigate each variable in a univariate analysis, but this may be better suited for a general linear model.

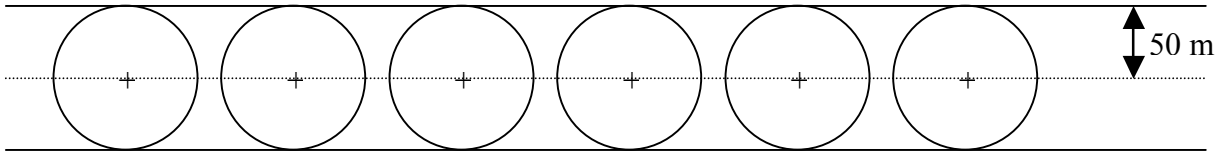
  Why are there 2 degrees of freedom (# levels – 1) in the ANOVA to test if carcass differed depending on whether the turbine was in a canyon? Either the independent variable is binomial (in a canyon or not) in which case there is 1 degrees of freedom or there were three "canyon categories" (yielding 2 degrees of freedom) that the authors did not articulate to the readers.

- p.43, Figure 2-13, p.44, Figure 2-14: The report of a strong effect of tower location within a string on the carcass distance is difficult to accept without careful analysis of the influence of the sampling method. The sampling method is described to some degree in Chapter 3, but it remains unclear how carcasses were associated with a particular tower within a string.

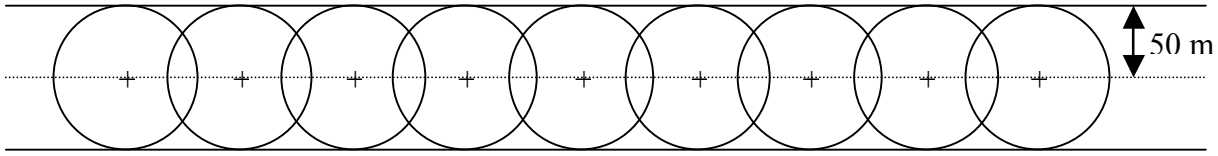
  For example, if the tower is not in a string (or if you prefer, a string of 1), then there is no confusion. Any carcass found within 50m of the tower is associated with that tower, and the search area would be $\pi \times 50^2 = 7854 m^2$. But for towers in strings, the tower spacing makes a difference. In the first sketch below, the towers are spaced more than 100m apart so that the area within 50m of each tower does not overlap with any other tower's area. (But looking forward to Figure 3-3 on page 51, will the search areas of a string then be very wide rectangles that include the spaces between the circles?) In the second sketch, the towers are less than 100m apart so there can be a lot of overlap in the 50m zone around each tower. Note that the end towers have greater area to themselves.

[7] Original review text before considering the Smallwood and Thelander response: " .. greatest influence on the regression. ~~If the 4 to 6 observations on the 43 m towers were removed, we suspect that neither of the two regressions would be statistically significant.~~ Even with the data..."

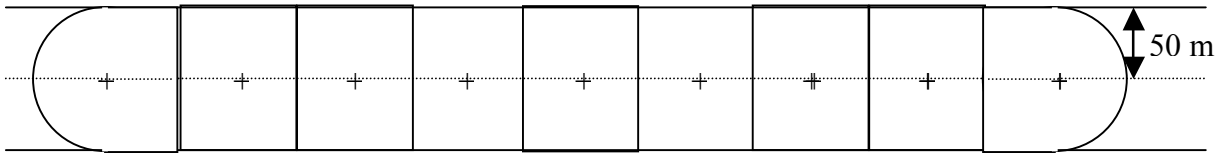
Case 1



Case 2



Based on Figure 3-3, a closer approximation to case 2 would be:



The towers not on the ends end up with rectangular search areas, where some of each rectangle is beyond 50m from a tower. On the other hand, the end towers in a string may have considerably larger sampling areas (depending on the tower spacing) and more at the further distances away from the tower.

For example, if the towers in the string were 50 m apart, then the search area for the towers in the internal string would be: height× width = (50m + 50m) × 50m = $5000m^2$. For the end towers, the area would be (half of a rectangle + half of a circle) = (50m + 50m) × 25m + 0.5× π× $(50m)^2$ = $2500m^2$ + $3927m^2$ = $6427m^2$ which would be 29% larger than for the internal towers.

So the distance between towers in a string is important.

- p.43, Figure 2-12: The authors show standard error plots of carcass distance by the different wind turbine types. Box plots would do a more adequate job of showing the spread of the data and inform the reader of potential biases in the study with regards to various wind turbine models. Specifically, box plots would show if distance of carcass (beyond 50m) would result in reduced carcass count for a particular wind turbine model. A "mean and 2 standard error" plot is designed to show the reader the range where the true mean is likely to be. With this study, however, we are more interested in the range and general distribution

of where the carcasses are to be found rather than what would be the long term average distance of where carcasses are to be found.

In addition, for large bodied birds, 50.0% of the carcasses are associated with KCS-56 turbines, 34.1% with Bonus, and 6.1% with Micron, totally 90.2% of just 3 of the 10 turbine types. Similarly for the small bodied birds, 83.6% of the total carcasses were found at the same 3 of 10 turbine types. How many turbines of each type are there? Is this disproportion chance or pattern?

Given the happenstance data collection on carcasses beyond 50 m, the inclusion of the beyond 50 m data in the analysis is inappropriate.

- p.44, Figure 2-13: Regarding distance of carcass from wind turbine for "end", "gap", and "interior" turbines and their analysis (p.44, par.1) could suggest that carcasses tossed far from one turbine could be attributed to the turbine to which it landed closest too. This is acknowledged in the discussion (p.45, par.3). Are all wind turbines in a string alike?

- p.45, par.1: The authors state that they found 15.3% of the large bird carcasses and 9.5% of the small bird carcasses outside of their 50m search radius. It is not surprising that only small percentages of the birds were discovered beyond 50m since the search effort in that region was happenstance. It is not stated whether these carcasses were found during the observers' systematic searches or while the observers were walking to the area where a systematic search would be done.[8]

- p.45, par.3: They state that extending their search radius to 100m would include 94% of the large bird carcasses[9] *they found, but this figure does not illuminate the number of carcasses that may still be missed beyond 50, because their rigorous searches terminated at that distance.* There is a well-established body of theory for estimating density of animals (or in this case, carcasses) using the distance to each detection and modeling probability of detection as a declining function of distance. There are computer programs (e.g., DISTANCE) for this sort of thing. These programs could essentially estimate the number of carcasses that were overlooked to yield a more unbiased and accurate estimate of carcass density.

[8] Original review text before considering the Smallwood and Thelander response: "... where a systematic search would be done. ~~How can you discover carcasses if you do not search for them?~~"

[9] Original review text before considering the Smallwood and Thelander response: "... large bird carcasses~~, an unsupported figure~~. There is a well-established body of theory..."

- p.45, Table 2-2: [10] The effects listed for Flowind and KVS-33 turbines are based on very small sample sizes (10 and 4, respectively) and also include happenstance discoveries beyond 50 m which further distorts the intervals. The reported effect could very well be spurious.

Chapter 3: Bird Mortality in the APWRA

- p.46, par.3: The authors propose that impact of the APWRA can be measured one of two ways: (1) number of fatalities per megawatt per year or (2) number of fatalities relative to the natural mortality and recruitment rates. They choose the fatalities per megawatt because it treats a certain number of fatalities as the "cost" of producing a megawatt. The other method evaluates the long term affect on the bird population; however, some of the needed demographic variables for such a measure are logistically unreasonable to estimate and beyond the scope of this project. Other authors use fatalities per turbine per year (p.46, par.4).

  It is more an issue of policy that determines which measurement is more helpful. Although not unreasonable, fatalities per megawatt per year ignores the total number of fatalities. Total number of fatalities is an important measure that shows, at least in part, an impact on the bird populations even if you do not know the demographic conditions of the species. Fatalities per megawatt per year is a good measurement if you are trying to minimize fatalities while producing a certain amount of energy. Fatalities per wind turbine would only be helpful if you are trying to minimize the number of fatalities for a fixed number of wind turbines regardless of energy output – something only reasonable if wind turbine models all had the same energy output.

  Another issue that needs to be looked at is how often, when operating, is a wind turbine actually achieving its rated energy output? Given that a wind turbine is operating, the distribution of time operating at various output levels will likely differ among different models. So would a higher rated wind turbine be more frequently operating at sub-maximum energy output levels than a smaller wind turbine, although being a risk to birds for as many hours as the smaller turbine? This would not be represented by the megawatt per year metric. The authors do briefly mention the lack of data regarding this issue on p.347, par.1 of Chapter 9.

- p.47, par.5: The authors sampled 1,526 wind turbines (182 strings) for 4.5 years and another 2,548 wind turbines (380 strings) sampled for about 6 months (November through May) because of access issues. Although this is about 75% of the wind turbines in the APWRA, the authors do not say how they decided

[10] This original review point began with the following paragraph before considering the Smallwood and Thelander response:. "~~This table summarizes the conclusions reached in this chapter about the distances of carcasses from towers. The relationships between distance and tower height are heavily influenced by a few observations on the tallest towers and in any case, the relationships are not substantial and only statistically significant in the most narrow technical sense given the r^2 values of 1%.~~"

which turbines to survey. Did they survey every wind turbine or string for which they had access? Did they use a sample of convenience, simple random sampling, or systematic sampling? The essential question is: Can the surveyed wind turbines be considered representative of the entire population of wind turbines or at least representative of the wind turbines for which they had access?

The short duration of sampling for the second set was the result of delayed access to the turbines from the owners. Although the first set includes fewer turbines and strings, it provides the primary and superior data set because of the repeated observations, the seasons sampled, and the increased duration. The limited duration of sampling, the lack of replication, and the restricted seasons sampled greatly reduces the value of the second set. Unfortunately, the analyses do not distinguish between the two sets.

- p.48, par.4: Was there any concern about whether severed body parts from one mutilated bird (wind turbine or scavenger caused) could have indicated more than one fatality?

- p.48, par.1 and p.49, par.2: The authors write, "...we recently found that 85%-88% of the carcasses occurred within 50m of the wind towers." The absence of any described systematic method of how they searched beyond 50m makes this estimate questionable. The authors then write the following:

> "Searcher detection and scavenger removal rates were not studied, because it had already been established that mortality in the APWRA is much greater than experienced at other wind energy generating facilities. We were unconcerned with the underestimating mortality, and in fact we acknowledge that we did so. We were more concerned with learning the factors related to fatalities so we can recommend solutions to the wind turbine-caused bird mortality problem. Thus, we put our energy into finding bird carcasses rather than estimating how many birds we were missing due to variation in physiographic conditions, scavenging, searcher biases, or other actions that may have resulted in carcasses being removed." (p.49, par.2)

With this statement, readers must treat all bird mortality estimates as relative estimates and not as the exact counts or unbiased estimates. Regardless, the authors go ahead and attempt to come up with reasonable mortality estimates.

- p.49, par.1: What is the sampling element in use in this chapter? The authors "… express mortality as the number of fatalities per MW per year …" The total

number of fatalities observed on a string divided by the total rated power output from the string and divided by the total duration of sampling. This indicates that the sample size is the string, so that each string, not turbine, has an associated fatality rate. So sample sizes should be the number of strings visited, not turbines visited.

- p.51, par.1: The authors did not assess searcher detection rates in this study and selected to use literature values: 85% detection rate for raptors and 41% for non-raptors. Solely in this chapter, these detection rate values are used to correct the observed counts for deficiencies in detection. This seems reasonable, but why do the authors feel detection would be 50% less likely to discover a small raptor such as a kestrel than a similar sized non-raptor, such as a robin? (This same question applies to scavenging rates as well.)

  They estimated the number of carcasses that actually existed by dividing either by 0.85 (raptors) or 0.41 (non-raptors). These calculations were equally applied to carcasses was found within or beyond the 50m search radius. This seems unreasonable to treat the beyond-50m carcasses the same as within-50m carcasses because carcasses beyond 50m were discovered by happenstance. The fraction missed beyond 50m could be much larger than their estimate.

- p.51, par.2 – p.52, par.2: The authors used scavenger removal rates and detection rates estimated in other studies to produce bird mortality estimates (p.51, par.1). A bothersome aspect of the authors' report is that they adjust the scavenger removal rates and detection rates from the other studies to rates that they believe better describe the APWRA and the time between their surveys without giving any anecdotal or empirical evidence of why they chose the numbers they did. Adding 10% to the scavenger removal rates of Erickson et al. (2003) to account for the authors' longer interval between searches appears arbitrary (p.51, par.2). Furthermore, without any support of data or other evidence the authors add (p.52, par.1), "Based on our experiences with raptor carcasses in the APWRA, we did not believe that these scavenger removal rates were accurate for raptors, and we halved the removal rate estimates reported by Erickson et al. (2003)." Underestimating scavenger removal rate will result in underestimating mortality.

  There is an error in their calculations for "halving" of the raptor removal rate. If s is the scavenging rate, the authors estimate the pre-scavenged carcass number by dividing the number of carcasses available after scavenging by $(1-s)$. After "halving" the scavenger rate, the authors simply divided by $2\times(1-s)$ while they should have divided by $1-\frac{s}{2}$. Their method reduced the scavenging rate by more than half and results in mortality estimates that are biased downward.

For example, the scavenger removal rate for carcasses of large-bodied species is 68.6% (p.51, par.2) thus the proportion of carcasses after scavenging to be found is $1 - 0.686 = 0.414$; therefore,

$$\text{Pre-scavenged number of carcasses} \times 0.414 = \text{Number of carcasses after scavenging}$$

.

So to calculate the pre-scavenged number of carcasses from the number of carcasses available to be found after scavenging, we divide by 0.414:

$$\text{Pre-scavenged number of carcasses} = \frac{\text{Number of carcasses after scavenging}}{0.414} \quad .$$

The authors halve the scavenging rate by doubling the denominator, thus 0.414 becomes 0.828. This, however, is different than halving the 68.6% down to 34.3% which would give an estimate of the pre-scavenged number of carcasses to be:

$$= \frac{\text{Number of carcasses after scavenging}}{(1 - \frac{1}{2} \times 0.686)}$$
$$= \frac{\text{Number of carcasses after scavenging}}{(1 - 0.343)} .$$
$$= \frac{\text{Number of carcasses after scavenging}}{0.657}$$

This does not equal the authors' pre-scavenged calculation of $\frac{\text{Number of carcasses after scavenging}}{0.828}$. Consequently the authors are more than halving the scavenger rate.

The combination of these various corrections results in an estimate of overall mortality that is, at best, rough and imprecise and, at worst, seriously biased (likely downward). *Inadequate*[11] consideration is given to these ad hoc corrections in evaluating the uncertainty in the mortality rate estimates provided later in this chapter.

As a last note here, the authors should make their calculations more clear to the reader. Erickson et al. 2003 provides a good template.

- p.52, par.2: The authors are correct in stating that their "mortality estimates might be conservative" because of removal of carcasses by people not involved in the

[11] Original review text before considering the Smallwood and Thelander response: "~~No~~ consideration..."

authors' study and they provide some anecdotal evidence. The authors do not account for such carcass removal.

- p.52, par.3: The authors state that, of the 1162 carcasses whose fatality was attributed to the wind turbines, 198 were more than 90 days old. Table 3.1 on pp. 64 and 65 counts fatalities as Type A (both fresh and old) and Type B (fresh; used to estimate mortality). The difference between Type A and Type B should be the number of carcasses older than 90 days. In fact the difference is 1162 – 923 = 239 which is larger than the 198 reported on p. 52. What happened to the other 41? Bats account for some, but not all.

- p.52, par. 4 and p.53, Figure 3-4: The authors state that the frequency distributions shown in Figure 3-4 are "at the string level of analysis". The caption for Figure 3-4 should reflect that the figure shows the frequency of strings with various levels of estimated mortality rates.

  It is striking that at 270 of the 562 strings searched, or 48%, no carcasses were found. A useful analysis would have been to compare the group of strings with zero fatalities to those with observed fatalities.

  Both parts of Figure 3-4 include what appears to be a truncated normal distribution. This is inappropriate since the observed distribution is quite unlike a normal curve, more closely resembling an exponential or Poisson distribution. The normal curves should be removed.

- p.52, par.5 and p.64, par.1: The authors make statements about inter-annual mortality variation for different species and types of birds at wind turbines sampled for all four years. It is assumed, but not stated, that ANOVA and LSD are used. The multiple categories of birds species/type being tested for inter-annual mortality variation makes the chance of at least one Type I error likely.

- p.52, par.5 and p.68, Tables 3-3 and 3-4: The statement about the mortality of burrowing owls based on the strings studied for 4 years vs. just 1 year refers to the right columns of Table 3-4. We suspect this should be Table 3-3.

- p.59, Figure 3-15: Year effects on mortality rate are confounded by location, as evidenced by this figure.

- pp.54-58, Figures 3-5 through 3-14: It seems as though the 95% confidence intervals in these figures were determined based on the string-based mortality rate estimates using Student's t distribution. Then it would be appropriate to provide

the sample size for each year and not just the aggregate for all 4 years. (Or was it a sample size of 160 for the 1-year strings and 62 for the 4-year strings?)

How was the confidence interval computed for 2001-2002 in Figure 3-9? It appears that the estimate is zero and the C.I. has zero width. How is this possible? Were there no barn owls killed in the 62 strings in 2001-2002? *If so, then the point should not include a confidence interval.*

- p.70, Table 3-9: To this point in this chapter, the analysis has been string based. This table refers to 1526 turbines in the first set and the 2548 turbines in the second set. The columns give the mean and standard error among strings, not turbines. What was the sample size used for each of the mean and standard error calculations? Is it number of turbines or number of strings? Are these sample sizes taken to be the same for all species or groups

  It would be useful to compare these results to the corresponding median values. It would be interesting to know how many of the median mortality estimates would be zero? Even for the shorter duration second set, 12 of the 30 (40%) species mean mortality rates are zero.

- pp.70-75, Tables 3-9 through 3-12: The authors should better explain the calculations used to produce these tables. An example using real data would be helpful.

- p.76, par.2: The authors mention high mortality estimates in the SeaWest-owned portion of the APWRA, but the Results (Section 3.3) did not articulate about spatial or owner differences in mortality rates.

Chapter 4: Impacts to Birds Caused by Wind Energy Generation

- p.78, par.3: The authors assume a 50% miss rate outside of their 50m search radius (p.78, par.3). This statement conflicts with their Chapter 2 methods (p.51, par.1) where they said the detection rate within 50m was the same as beyond 50m. Thus in Chapter 2 they used detection rates for beyond 50m of 85% (raptors) and 41% (non-raptors). A 50% detection rate beyond 50m for non-raptors would suggest a greater detection rate beyond 50m than within 50m, obviously not sensible. More reasonable detection rates would be 42.5% (raptors) and 20.5% (non-raptors) beyond 50m (i.e., half the detection rate as within the more thoroughly searched 50m).

- p.78, par.4: The authors present findings from point count surveys although they have not yet discussed the methods with the readers.

- In general, Chapter 4 does not adequately portray that the mortality estimates at APWRA from this report are likely biased low – perhaps severely. This bias comes about because: (1) detection rates for carcasses beyond 50 m could easily be well below the values used in analyses; (2) scavenging rates could easily be higher than used in analyses (because search intervals were longer for this study than in the studies from which values were obtained); and (3) scavenging rates of raptors were arbitrarily cut in half from reported scavenge rates.

Chapter 5: Range Management and Ecological Relationships in the APWRA

- In general, the authors present the reader with a blizzard of one-way ANOVA and LSD statistical tests looking at an almost endless number of variables. Having so many variables inspected individually, leaves the study highly vulnerable to Type I errors, confounding variables and difficult to interpret findings. A multivariate approach would help the authors develop a more thoughtful, concise analysis that can help control for confounding variables.

- p.91, par.3: "Vegetation height ... was 18% greater ... where rodenticides were intermittently deployed...," the authors report with a mean difference from intense rodenticide use of 4.28cm. The magnitude of 4.28cm is more meaningful if the mean heights of the grasses are also provided. It could be 1cm vs. 5.28 or 11cm vs. 15.28 which could understandably have different ecological impacts.

- p.100: The authors indicate that the index of cottontail rabbit abundance was higher on Enertech towers, on plateau slope combinations, and on southwest slopes. Were Enertech towers especially common on southwest slopes relative to other tower types? These questions are difficult to answer because they require the reader to extract information presented for other purposes elsewhere in the report. By running multivariate analyses (which may require simplifying or reducing variables – in itself a good thing), then the association between a given predictor variable and the response variable can be measured while statistically accounting for confounding variables. This is a recurring limitation of the study.

- p.103, Table 5-20. This is an example of where the authors should interpret the meaning of the analyses while paying attention to the magnitude of differences. Furthermore, the metric "cottontail abundance" is never defined. In Table 5-20 cottontail abundance is compared between "some lateral edge" and "other edge conditions" with a statistically significant "Mean difference (cm) on grass

transect" of 0.18. What does that 0.18cm represent? Is that a small biological magnitude that ends up being statistically significant because of the very large sample size of 1327?

- p.108, par.4: The authors make quick mention that, "Some of these relationships might be confounded with other variables." This is an understatement and a recurring limitation of the study. Multivariate analyses could help control for some of these confounding variables.

Chapter 6: Distribution and Abundance of Fossorial Animal Burrows in the APWRA and the Effects of Rodent Control on Bird Mortality

- p.111, par. 4: "Most wind turbine strings were selected arbitrarily, to represent a wide range of raptor mortality recorded during our fatality searches, as well as to represent a variety of physiographic conditions and levels of rodent control," the authors write. A more rigorous method of selection should have been used, such as stratified sampling. The objectiveness and unbiasedness of "arbitrary" sampling is always questionable.

- p.112, par.4 and p.114, par.5: The method of estimating degree of clustering at wind turbines using the slope from least squares linear regression is unclear (p.112, par.4). Is "corresponding search areas" the distance from the wind turbine? It then seems that the authors disregard this "regression-slope" method (p.114, par.5) for the "observe*d*-divided-by-expected" approach. Having this "regression-slope" method discussed is confusing if it is not to be used.

- p.112, par.6: The authors mention that they learned *post hoc* about rodent control. Although likely beyond the duties of the authors, the effectiveness of rodenticides to reduce raptor mortality could be better explored in the future via a carefully planned experiment.

- p.149, par.5 and p.164, par.1 and Figures 6-45 and 6-46: The simple linear regressions used to investigate association between raptor mortality and ground squirrel burrow systems are very questionable (Figures 6-45 and 6-46). The authors discuss the significance of these scatter plots (p.149, par.5 and p.164, par.1). Some of these conclusions and "significant" *P*-values are based on sample sizes of 3 (no rodent control) and 5 (intense rodent control) – it is *outside the realm of professional practice*[12] to base inferences from just 3 or 5 data points.

[12] Original review text before considering the Smallwood and Thelander response: " ... it is ~~foolish~~ to base inferences..."

Furthermore, leverage of an individual point affects all three levels of rodent control and the assumption of homogeneous error is ignored.

- pp. 164-172, Tables 6-2 through 6-11: These tables aggregate the density of burrows into categories and then total the number of bird kills for each of the three categories. It is not clear how the authors decided to define each category and information is lost by categorizing continuous data. A dot plot or histogram of the burrow densities for where carcasses were found beside a second plot of burrow densities for where carcasses were not found would have been more informative.

- Discussion, pp.172-178: The authors make good points in the Discussion regarding the negative and/or inconsistent impacts of rodent control measures, and their case is strong, we believe. They offer the caveat that, "Intense rodent control was associated with fewer golden eagle fatalities in areas of intense rodent control, but the association is not strong enough to warrant its continued use" (p.178, par.2). We think that statement is giving the rodent control measure more causal credit than it deserves. In fact, the *P*-value for the ANOVA test of golden eagle mortality rate across the three rodent control intensity levels is statistically insignificant at 0.9 (p. 172, Table 6-12). While the mean mortality estimate is slightly lower in magnitude for the intense control category, the variance is very large, and we thus have no confidence this difference is "biologically real." One could just as easily claim that, "mortality rates among rodent control intensity were statistically indistinguishable."

Chapter 7: Bird Fatality Associations and Predictive Models for the APWRA

- p.182, par.5: The authors define four seasons, but the length of the seasons are very different: spring is 92 days, summer is 117 days, fall is only 51 days, and winter is 105 days. Summer is 2.3 times as long as the fall. What is the justification for these definitions? The authors also give no explanation of how they decide "number of days since death" when a carcass is discovered.

- p.182, par.7: Although Table 1-1 does summarize the attributes of the wind turbines in the sample, it does not state the frequency of each type in the sample and the population.

- p.183, line 7: The authors need to be careful and consistent as to how they show their mathematics. They most often, but not always, use more elementary notation such as $A \div B$ instead of $\frac{A}{B}$. On the 7th line of page 183, they define "the

window of opportunity" as $\text{Window} = \text{C} \div \text{T} \cdot \text{B}$. This is equivalent to $\frac{C \cdot B}{T}$, but the equation is more sensible as $\frac{C}{T \cdot B}$, which we believe is what the authors meant. The authors should employ the use of an equation editor, like that used in Microsoft Word.

- p.183, par.2: For purposes of computing how quickly a bird clears the rotor plane, how thick is the plane? What flight speed would be required to clear the rotor plane in the allotted time?

- p.183, par.5: The tower height is defined as the distance the rotor is above the ground. Can we assume that this is the center of the rotor?

- p.184, par.1: The incidence of rock piles was reduced to a limited number of categories. Did the authors intend the categories to be: a) none, b) less than or equal to 0.25 piles per turbine, or c) greater than 0.25 piles per turbine?

- p.184, par.2: The authors employ a 40 m radius around each turbine instead of the 50 m radius stated earlier. What is the reason to redefine the sampling zone now?

- p. 184, par.4: Did the authors test the assumptions of the statistical tests (e.g., homogeneity of variances or statistical independence and normality of residuals) applied in this or any other chapter? What objectives are the authors trying to meet in reporting "weak and non-significant correlations"? How can the measures of effect*, statistically or biologically,* be meaningful if the confidence interval for the magnitude of the effect includes zero? *A nonsignificant result would imply a confidence interval that includes zero.*

- p.184, par.5: For regressions, the authors have chosen to include the RMSE to provide a measure of the "precision of the data relative to the regression line". By RMSE, we assume that the authors mean:

$$RMSE = \sqrt{\frac{Sum - of - squared - residuals}{samplesize}}$$

A more appropriate estimator for precision*, for either simple or multiple regression,* would have been the standard error of the estimates (SEE) or:

$$SEE = \sqrt{\frac{Sum - of - squared - residuals}{samplesize - \#of\ parameters}}$$

- p.184, par.6: Although this is a non-manipulative study and the existing towers, turbines, topography, etc. as well as permission for access does limit the range of choice, it is still possible to carefully select the areas of study to provide the contrasts and comparisons of interest.

- p.185, par.1: Is the term "efficient" used here in the technical sense from statistics?

- p.185, par.2: The authors discuss the 5% significance level used in the subsequent tests and the 10% level that they interpreted as indicating "trends worthy of further research". Given the immense number of univariate hypothesis tests reported in the subsequent pages, the authors should have discussed the risks of Type I errors (false positives) associated with conducting hundreds of tests.

  The total number of chi-square tests presented just in Tables 7-1, 7-2, and 7-3 is 528 (ignoring the many more chi-square tests presented in Appendices B & C). The chief disadvantage of this approach is that Type I and Type II (false negative) error rates are inversely related, creating no clear optimization. One could argue that Bonferroni adjustments are necessary to guard against very high experiment-wise Type I error stemming from so many tests. Using Bonferroni adjustments, the experiment-wise alpha (level of significance) value "should" be set as: $\alpha_{adj} = 1 - (1 - \alpha)^{\frac{1}{n}}$; in this case $\alpha_{adj} = 1 - (0.95)^{\frac{1}{528}} = 0.000097$ for a modified Bonferroni adjustment as proposed by Shafer (Shaffer, J. P. "Multiple Hypothesis Testing." *Ann. Rev. Psych.* **46**, 561-584, 1995.)

  But if the authors bring the experiment wise alpha value this low, the Type II error rate gets unacceptably high, especially for work designed to measure environmental impact. That is, the probability of the analysis suggesting no impact when in fact there is one becomes unacceptably high. This problem further underscores the value of a smaller number of multivariate tests, as we have suggested elsewhere.

- p.185, par.3: The uses of chi-square tests "for association" are described. The chi-square tests used by the authors are more commonly described as chi-square tests for "goodness-of-fit" where they are testing whether it is plausible that the observed counts across the categories came from a uniform distribution (each category is equally likely). Although statistically legitimate, such methods fail to control for other variables, leaving the study vulnerable to confounding variables.

Why not use a general linear model, logistic (yes/no data) or Poisson (counts data) regression, discriminant analysis, or at least a log-linear analysis?

- p.186, par.3: The authors rationalize that relative search effort can be calculated as, $N_t \times R \times Y$, where N_t is the number of wind turbines in a string, R is the mean rotor swept area in m^2, and Y is the number of years the string is searched. This decision is based on Figure 7-1. It is a loose association between the relative search effort and number of fresh bird carcasses found. From this, they assume that mean rotor swept area is proportional to the number of carcasses – a circular argument since that is what they are supposed to be investigating. Keep in mind that the swept area is proportional to the squared radius of a wind turbine (Area=$\pi \times r^2$), thus the "search effort" at a wind turbine with a 3m blade will be four times as much as at a wind turbine with a 1.5m blade (half the size) even if they physically searched the surrounding grounds equally. Thus the wind turbine with a 3m blade will have to kill four times as many birds to have the same rate of mortality as the 1.5m blade wind turbine, ignoring megawatt output. In Appendix A, the authors do show a positive relationship between megawatt output of a turbine and mortality. Perhaps the authors are trying to copy epidemiology studies which use "people years" when calculating risks for cancer; e.g., following 100 people for 5 years is equivalent to following 250 people for 2 years. Here this would correspond to "turbine years". It is a strong assumption to say that the variable "rotor swept area" is just as important as the variables "time" or "number of wind turbines" with regards to the number of expected bird carcasses.

- p.186, par.4 and p.187, Figures 7-1 A & B: Figure 7-1A presents the relationship between the number of birds recently killed at turbine strings and the measure of search effort used. [13] Which of the variables account for the observed variation in the search effort: the number of turbines in the string, the mean rotor swept area, or the number of years of searching?

  The authors suggest that Figure 7-1B illustrates an inverse power relationship between fatality rates and search effort. It would be more informative to plot the data shown on a log-log plot, which would more conveniently indicate if the relationship was in fact an inverse power relationship. It appears, however, that there may be many observations with fatality rates of exactly zero, but it is difficult to tell since the vertical axis does not show a zero.

[13] Original review text before considering the Smallwood and Thelander response: "...search effort used. ~~Of the 472 data points, only 32 or so exceed 10,000 m^2-yr of search effort and only 2 of the 472 exceeds 30,000. Consequently, these extreme values of the total dataset have the principal influence on the regression results.~~ Which of the variables account..."

Figure A4 (p. A-8) suggestions a mechanism that would produce the relationship suggested for Figure 7-1B. This indicates that the sampling approach yields stable estimates only after longer periods of search, which should be discussed here.

- *14*

- p.189, par.2: So now the sampling element is the wind turbine and no longer the string. What fraction of the total population of wind turbines does this sample of turbine models represent? It is important to the reader to know if these sampled wind turbines are representative of the APWRA population of wind turbines.

- p.190, Figure 7-2: The figure shows that the authors' study is essentially a study of KCS-33 and Bonus wind turbines. Furthermore, the "effort" for Bonus wind turbines is almost three times that of the number of Bonus wind turbines studied. Is that a result of the "relative effort" definition and that Bonus wind turbines' rotor sweep area is three times that of most other turbine models?

- p.189, par.8 and p.202, Figure 7-18: Based on the authors' definitions of seasons, fall is the shortest season (51 days) and so would be expected to have less sampling effort. Given the length of the seasons and assuming a uniform distribution of sampling times throughout the year, we would expect 25% of the observations in the spring, 32.1% in the summer, 14.0% in the fall, and 28.8% in the winter. Comparing this to the bar heights in Figure 7-18, the sampling effort is higher than expected in the spring, lower in the summer, higher in the fall, and on target in the winter. Is this a result of their sampling effort definition? It is not clear.

- p.192, Figure 7-4: Why is effort so many times greater for the wind turbines with 2141 rotor plane swept per second?

- *15*

[14] The original review's point was removed before considering the Smallwood and Thelander response:: "~~p.188, par.2: "Positive values express the percent of total fatalities likely killed at wind turbines due to the attribute associated with the value..." The use of the word 'due' implies causality, although at best they can only claim 'association'.~~"

[15] The original review's point was removed before considering the Smallwood and Thelander response:: "~~pp.193 and 194, Figures 7-5 and 7-6: These figures show scatter plots where an outer single point has high leverage (influence). Conclusions are essentially being determined by the one point furthest to the right.~~"

- p.199, Figure 7-14 through p.201, Figure 7-16: Why are the bin widths increased in going from graph A to graph B for each set of graphs? In graph B of each pair of graphs, the bin widths are not equal.

- p.203, Table 7-1: The dangers of multiple hypothesis testing arise in Table 7-1 when 204 chi-square tests are performed. (This is repeated again in Tables 7-2 and 7-3.) This can be kindly called "data exploration" or criticized as a "data dredging". Regardless, with 204 statistical tests, if all data were a result of a uniform distribution across each category, researcher error or biased post-hoc categorization did not cause any non-uniform distribution, and each test were independent of one another, you should expect 5% of the tests to give p-values less than 0.05. So there is a high chance of Type I errors when so many tests are performed. Also many variables may be correlated, such as "tower height" and "high reach of blades". So if a test was significant for "tower height" you should expect it to also be significant for "high reach of blades". In addition, a more clear explanation is needed as to why some variables such as "rodent control" and "Slope aspect" are tested twice.

  There are methods to help reduce the problems of multiple testing, such as Bonferonni corrections that make the p-value for declaring a "statistically significant result" much less than 0.05 for each test. This makes the overall chance of a Type I error only 5% if all tests were actually not significant. The problem with such adjustments is that the statistical power then decreases for each test opening the door for Type II errors thus making the researchers miss important variables. The authors should take a more selective and thoughtful approach to investigating the variables and use generalized linear models or multiple regression. These more advanced methods would help reduce some confounding by allowing the authors to control for other variables when testing another. The authors did, however, state that they only used the predictive model for variables that were statistically significant and showed gradients along a continuum (p.188, par.3).

  Furthermore, what are the sample sizes for each of these chi-square tests? A large sample size can produce very small p-values (very high statistical significance) even though the magnitude of difference from the uniform distribution is minimal; i.e., lacking biological significance. When the authors discuss the finding from the chi-square tests, they report something along the line of, "Wind turbines with variable X killed disproportionately more birds of species Y." What magnitude is implied by "disproportionately"? With a large enough sample size, it could be a biologically insignificant increase that is likely just a result of confounding. This issue of magnitude is addressed in Table 7-5 (p.215), but the percent magnitudes still need to be put side-by-side with real numbers to make them more meaningful.

- pp.207-209, Figures 7-19 through 7-21: There appears to be considerable spatial clustering of the golden eagle, red-tailed hawk, and burrowing owl fatalities. The variation in duration of study does not coincide with the clusters. Similar spatial clusters appear in all three figures. There is no discussion of *these figures*[16] in *this* narrative. Are these clusters the result of turbine type clustering, variation in elevation, concentration of avian habitat, or some other factors?

- pp.210-219, Tables 7-4 through 7-7: Percentage increases in mortality are listed for various species in association with 12 factors. Confidence intervals should be provided for each of these percentage values so that the precision of the estimated effect can be evaluated. How many of these confidence intervals would include zero, indicating that the magnitude of the effect might plausibly be zero?

- p.219, par.1 and pp.220-221, Figures 7-22 and 7-23: The authors note the seasons with relatively higher fatalities than expected but neglect to point out the seasons with unusually lower fatalities than expected. Specifically, the red-tailed hawk, American kestrel, and burrowing owl all show much lower fatalities than expected in the spring. Why would this be true? Similarly, there were no fatalities of mallards in the fall. Why would this be so?

p.222, par.2: "The empirical models developed were tested only against the database of the 4,074 wind turbines from which the data were obtained for model development," state the authors. Testing the quality of a statistical model on the same dataset from which it was developed is bad practice. The selected model may fit that specific dataset well, but not be robust enough to predict outcomes well from a similar but different dataset. Some statisticians, for example, will randomly set some fraction of the original data to the side (test set), fit a model on the remaining data (learning set) and then see how will it predicts the data that had been set aside. This is repeated until all data have been set aside once in the test set. Once a good model has been determined, it is fit to the entire dataset. This concept is much-addressed in the ecological statistical literature (e.g., Fielding and Bell 1997, Boyce et al. 2002, Knightes and Cyterski 2005), and there are numerous analytical approaches to minimize the circularity without requiring the collection of new independent data. The authors need to address these issues.

Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38-49.

Boyce, M. S., P. R. Vernier, S. E. Nielsen, and F. K. A. Schmiegelow. 2002. Evaluating resource selection functions. Ecological Modelling 157:281-300.

Knightes C. D. and M. Cyterski M. 2005. Evaluating predictive errors of a complex environmental model using a general linear model and least square means. Ecological Modelling 186:366-374.

[16] Original review text before considering the Smallwood and Thelander response: "There is no discussion of ~~this~~ in ~~the~~ narrative."

- p.222, par.3: This argument is independent of any observations made by the authors. It represents circular reasoning. It argues that if the model is correctly predicting which turbines are relatively more dangerous, then the reason no bird fatalities were found at most of these dangerous turbines is just that we did not look long enough. This might be true but this work can neither support nor refute it.

- p.223, Table 7-8: The authors have so far only conducted univariate chi-square hypothesis tests. They now seek to combine the results in an ad hoc fashion into a model which amounts to a scoring system. If the authors want to develop a multivariate model, they should apply appropriate methods such as logistic or Poisson regression.

- p.224, Table 7-10: The authors' interpretation of the results presented in this table is unusual. They group the observations by the results (e.g., 0, 1, 2, 3, etc. fatalities) and compare the fractions that were predicted to be "more dangerous" and "less dangerous". This is a backwards approach to evaluating the predictive model. The observations should be grouped by the predictions (not the results) and the percentages of each group that experienced fatalities should be compared.

  For example, using the golden eagle data, we can assemble a 2 x 2 table of relative risk:

	Predict 0 fatalities	Predict ≥1 fatalities
Observed 0 fatalities	2007	2014
Observed ≥ 1 fatalities	10	43
Total	2017	2057
% with fatalities	0.5%	2.1%

  So although the turbines predicted to be more dangerous were about 4 times more likely to experience fatalities than the turbines predicted to be less dangerous, 97.9% of those predicted to be more dangerous experienced zero fatalities. On p.222, par.3, the authors argued that this large rate of false positives is attributable to the short duration of sampling. If so, then the turbines studied for 4 or more years should show a stronger response. Is this effect stronger for the turbines studied for longer periods?

- p.226, p.229, p.231, p.235, Figures 7-24, 7-26, 7-28, 7-30: In the A part of each of these figures, the authors have again grouped the observations by the results and not the predictions. Since they are attempting to evaluate the quality of the predictions, their approach is inappropriate. Like residual plots for logistic regression, the observations should be grouped by prediction (ranges of the scores) and the fraction of turbines experiencing fatalities should be compared among the prediction groups.

- p.242, par.1: The authors state that they "… were unable to account for interactions effects between independent variables." If more appropriate tools had been applied to the model development, investigation of interaction effects would have been straightforward.

- p.242, par.2: The authors claim that elimination of 20% of the turbines might reduce the mortality by 40%[17]. How was this determined?

- p.243, par.2: The authors state that the Bonus, Micon, and KVS-33 turbines are the most dangerous. How was this determined? It is likely the authors intended to include the KCS-56 instead of the KVS-33 based on the total bird fatalities reported in Table D-3. Is it possible that there are more fatalities for these turbines because there are more of them, not that they are more dangerous per unit?

- p.245, par.2: The authors state that wind turbines that are at the end of strings or are isolated kill more birds that wind turbines on the inside of strings. It is important to keep in mind that carcasses tossed far enough by a wind turbine that is on the inside of a string can be misattributed to either its left or right neighbor. Wind turbines at the end of a string can only have their kills misattributed to another wind turbine only if it tossed towards the string. Wind turbines that are isolated will not have any chance of getting their carcasses misattributed.

Chapter 8: Bird Behavior in the APWRA

- p.246, par.4: Biologists only collected bird behavior data from mid-October through mid-May. What about mid-May through September, especially since summer is when the winds are strong? Perhaps young prey or different types of prey are available more during certain months? Also, how were the 61 observation plots selected: randomly or by convenience?

- p.247, par.2.: The observation plots had a fixed radius of 300 m, so the term *variable distance circular point observations* is not really appropriate. Variable-radius plots are more commonly used in so-called "distance based sampling" in which the distance to *each* bird observation is used to estimate probability of detection as a means of calculating bird density (which is not the intent of the authors). The authors did assign birds to one of 3 distance categories (based on distance to turbine), but the furthest category was truncated at 300 m. As

[17]Original review text before considering the Smallwood and Thelander response: "...by ~~80~~%."

Reynolds (1980) states, "With the variable circular plot method no maximum distance restrictions are placed on any observation" (p.310). "Distance-based sampling" is a large sub-discipline within wildlife ecology and boasts a sizeable literature (see Volume 119 Issue 1 [2002] of *The Auk* for several recent papers on this subject), and while Reynolds et al. (1980) is a classic citation and influential in the development of current methods, it is not up-to-date with recognized methods.

- p.247, par.3: The authors state that the 61 observation plots were sampled 4 times each or "once every three to four weeks". How can the sampling cover 210 days and at the same time be once every 21 to 28 days? With one sampling at the start and one at the end, the interval between samplings would need to be about 70 days.

- p.250, Table 8-2: More explanation is needed to distinguish the types of flight behavior in Table 8-2. Contouring and surfing sound alike.

- p.251, par.1: The authors assume that, "the number of on-the-minute observations represented the same number of continuous minutes of the same activity." This is a standard assumption with conventional wildlife behavioral sampling, and is likely valid if sample sizes are large enough. This issue has been discussed extensively in the literature (see classic book by Martin and Bateson, 1993), the authors should make use of citations on the subject and defend that the assumption is valid. Also, they should identify their sampling technique within the conventional behavioral sampling lexicon – i.e., there are very standardized differences between focal animal sampling, scan sampling, and instantaneous sampling. The authors likely did the latter, but they should review these terms and identify which best describes their approach. (Martin, P. and P. Bateson. 1993. Measuring Behavior, An Introductory Guide. Cambridge Univ. Press, London, UK.)

- p.253, par.5: Chi-square tests are performed to test for disproportionate behavior under various conditions. Observations (data points) used in a chi-square test should be independent of one another. Having a single bird provide multiple observations through time removes that independence, thus invalidating the chi-square analysis. If a bird is soaring one minute, it is more likely to be soaring during the next minute. Even if a bird only contributed one observation; it could be recounted as a new bird if it disappeared for only 30 seconds (p.247, par.5).

- p.260, par.3 and p.260, Figure 8-9: The authors state that an asymptote for some behaviors is reached by about 9 minutes and for others by 20-27 minutes. It is not clear what asymptotes they are referring to. The vertical axis on Figure 8-9A

does not include zero, which exaggerates the magnitude of the change. Why did the frequency of behaviors increase with time? Does this suggest birds took some time to habituate to human presence (as suggested by Reynold et al. 1980 and others)? Or does it mean in took 8-30 minutes for observers to begin to fully "notice" (authors' term) behaviors in the observation plots? The term *special behaviors* is inadequately defined.

- p.256, par.5: The authors absolutely did not observe 855 minutes of flying; they recorded 855 incidences of flight among 3884 observations at minute intervals. There is a difference between these two. This is a problem with equating minutes of an activity with frequency of its observation at 1-minute intervals.

- p.256, par.6 and p.262, Figure 8-11: The authors state that Figure 8-11A shows the relationship between the number of flights through the rotor zone and the total number of flights observed during a session. What is the slope, r^2 value, or standard error estimate for the relationship? Is this a chance pattern? Regardless, it makes sense that if there are more incidences of flight, there will be more incidences of flight through the rotor zone. And if birds are perching – thus not flying – there will be fewer incidences of flight through the rotor zone.

- p.264, par.2: Were any bird collisions with turbine blades observed?

- p.265, Table 8-3: The table totals for the sum of minutes of flying (855) does not match the total of the column (828). Are there other raptor results not tabulated? Similarly the total provided for the sum of minutes perching column is 3029 but the column total is 2909. And the total given for the number of flights through the rotor zone (153) does not agree with the column total of 147.

  The turkey vulture, red-tailed hawk, and American kestrel account for 87% of the minutes flying and 90% of the flights through the rotor zone, but according to Table 3-1 they only account for 22.9% of the total turbine caused fatalities and for 58.1% of the total raptor fatalities cause by collisions. Why this great disparity?

- p.266, Table 8-4: In this table there are several behaviors or groups of behaviors that have zero recorded minutes of activity for all listed species and yet three other flight behaviors listed in Table 8-2 are not included (e.g., high soaring, mating, and land). Why were these omitted?

- p.267, Table 8-5: There is a discrepancy between the minutes perching for American kestrels between this table (1065) and Table 8-3 (1103).

- p.269, par.4: Many of the environmental variables may have coincidentally been correlated with when the birds were sighted. For example, "Golden eagles and American kestrels perched more often than expected by chance during cooler temperatures, which was also more or less when they flew more often." So were Golden eagles and American kestrels mostly observed during the cooler months? Would *such confounding*[18] also cause an association with certain seasonal types of wind? And how can they be perching more and flying more at the same time? Would not one increase while the other decreases?

- pp. 270-275, Tables 8-6 through 8-11: The authors have again conducted 132 univariate hypothesis tests without correcting for multiple comparisons.

- pp. 283-307, Tables 8-12 through 8-16: This time there are 792 simultaneous tests conducted without correction for multiple comparisons.

Chapter 9: Conclusions and Recommendations

- p. 339, par.3: The authors state that birds are disproportionately killed by wind turbines mounted on tubular towers. However, because of the tubular vs. lattice towers differ in many other respects (rotor length, tip speed, blade height, etc.), without examining effects of tubular vs. lattice towers while controlling for the other confounding variables via multivariate analysis, the univariate analyses are suspect.

- p.353, par.5: The authors state:

> "We also had little control over the application of sampling effort across the APWRA, and so the differential sampling effort we applied precluded multivariate statistical methods, which would have been useful for managing the shared variation among measured variables. These factors required us to rely on univariate tests."

The lack of management of shared variation among variables is indeed a major limitation of this study. But unrepresentative, incomplete sampling is a problem for univariate as well as multivariate analyses. There is no reason why the authors cannot employ more state-of-the art analytical tools to try to disentangle

[18]Original review text before considering the Smallwood and Thelander response: "Would ~~that~~ also cause..."

the multiple measured variables, with the strong caveat that the sampling was likely inadequate.